International Journal of Engineering, Management and Humanities (IJEMH)
Volume 3, Issue 5, pp: 39-47 www.ijemh.com

Classification Of Medical Prescription Using Nlp

DEBMALYA RAY

Date of Submission: 10-09-2022 Date of Acceptance: 25-09-2022

ABSTRACT

This data was scraped from mtsamples.com. The data contains sample medical transcriptions from various medical specialities. The idea is to use these transcriptions and classify them to the medical specialties based on various NLP techniques.

The extensive growth of data in the health domain has increased the utility of NLP in health. A vast amount of data in the form of text are generated by medical departments through medical prescriptions. We tried to perform different comparative analysis using classification algorithms and advanced techniques like BERT.

One key issue is that medical information is presented as free-form text and, therefore, requires a time commitment from clinicians to manually extract meaningful information. Natural language processing (NLP) methods can be used to extract relevant information and perform classification methods on it.

The BERT model has arisen to be popular in recent years. It can cope with NLP tasks such as supervised text classification with better feature engineering techniques. In this paper, we also tried to cover the concept of BERT and its application for this particular problem statement.

Keywords, medical, classification problem, machine learning, vectorization, feature engineering, NLP, BERT

I. INTRODUCTION

The Natural Language Processing (NLP) methodologies have flourished and lots of papers solving different tasks of the field, such as text classification, named entity recognition or summarization have been published. The search for a universal representation of text is at the heart of the automated processing of natural languages. Over the last few years, many new methodologies and feature engineering techniques have evolved which include stemming, lemmatization and vectorization.

This study focuses on extracting medical or clinical data and using scientific approaches to find out its medical specialties.

The machine learning models use the below data flow (**Figure 1**) as a part of the process to perform the necessary task. The train data identified has to be changed into its numerical form before feeding into the model. There are specific feature engineering techniques which are discussed in detail in the Research Methodology section to perform this task. Same steps as performed for the test data as well.

Data Flow Representation:

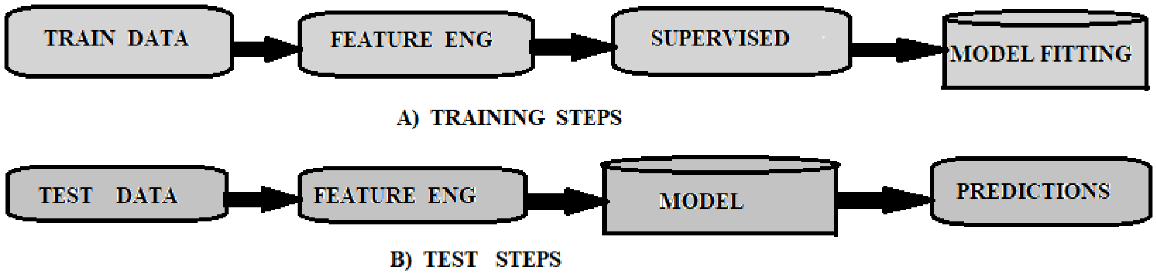


Figure 1

ML models used:
We did a comparative analysis with various ML models with their accuracy and find out the best model suitable for this problem statement. The algorithms used for this study are mentioned below:
a) Classification Algorithms
b) BERT

II. Problem Statement

Problem:
To classify medical specialties based on the transcription text collected as a part of the data preparation stage.

Background:
It is a challenging task to perform classification analysis based on medical data or clinical data collected from various sources. With the advancement in the health sector, it become important to collect clinical data, mostly in text format and perform the techniques applicable to the problem statement as discussed above. A few papers were already published that mentioned how we can efficiently use clinical data and help patients for treatment purposes.

Each of these studies used various metrics and techniques to perform such analysis. We are trying to drill into it and find out the best suitable techniques to categorize medical transcriptions into their specialties.

III. LITERATURE REVIEWS:

Understanding The Problem	Citation
Results show a high degree of accuracy in the extraction of drug-related information. Contrastingly, a much lower degree of accuracy is demonstrated in relation to auxiliary variables. In combination with state-of-the-art active learning paradigms, the performance of the model increases considerably.	Nemanja Vaci, Qiang Liu, Andrey Kormilitzin, Franco De Crescenzo, Ayse Kurtulmus, Jade Harvey, Bessie O’Dell, Simeon Innocent, Anneka Tomlinson, Andrea Cipriani, Alejo Nevado-Holgado "Natural language processing for structuring clinical text data on depression using UK-CRIS"
Advancement in pre-trained text encoders using models like BERT	Ian Tenney, Dipanjan Das, Ellie Pavlick "BERT Rediscovers the Classical NLP Pipeline"
The paper describes the mechanism of operation of this model, the main areas of its application to the tasks of text analytics, comparisons with similar models in each task, as well as a description of some proprietary models.	Koroteev M.V. "BERT: A Review of Applications in Natural Language Processing and Understanding"
Prescription medication (PM) misuse/abuse has emerged as a national crisis in the United States, and social media has been suggested as a potential resource for performing active monitoring. However, automating a social media-based monitoring system is challenging—requiring advanced natural language processing (NLP) and machine learning methods. In this paper, we describe the development and evaluation of automatic text classification models for detecting self-reports of PM abuse from Twitter.	Mohammed Ali Al‑ Garadi1*, Yuan‑ Chi Yang1, Haitao Cai2, Yucheng Ruan3, Karen O’Connor2, Gonzalez‑ Hernandez Graciela2, Jean Marie Perrone4 and Abeed Sarker1 "Text classification models for the automatic detection of nonmedical prescription medication use from social media"
The accuracy of medication data in electronic	Wei-Chun Lin, Jimmy S. Chen, Joel Kaluzny, Aiyin

health records (EHRs) is crucial for patient care and research, but many studies have shown that medication lists frequently contain errors. In contrast, physicians often pay more attention to clinical notes and record medication information in them. The medication information in notes may be used for medication reconciliation to improve the medication lists' accuracy. However, accurately extracting a patient's current medications from free-text narratives is challenging. In this study, we first explored the discrepancies between medication documentation in medication lists and progress notes for glaucoma patients by manually reviewing patients' charts.	Chen, Michael F. Chiang, Michelle R. Hribar "Extraction of Active Medications and Adherence Using Natural Language Processing for Glaucoma Patients"
To investigate community pharmacists' knowledge, attitudes, perceptions and habits about antibiotic dispensing without medical prescription in Spain.	Juan Vazquez-Lago, Cristian Gonzalez-Gonzalez, Maruxa Zapata-Cachafeiro, Paula Lopez-Vazquez, Margarita Taracido, Ana López, Adolfo Figueiras "Knowledge, attitudes, perceptions and habits towards antibiotics dispensed without medical prescription: a qualitative study of Spanish pharmacists"

IV. Aim and Objectives

Aim of the study:
To propose the best text classification technique and perform a comparative analysis with various machine learning models and advanced NLP so that a transcript can be categorized to its Specialty.

Based on the aim, we have created a set of objectives as mentioned below:

- To define the data frame suitable for the dataset
- To find out the best data cleansing and engineering techniques.
- To analyze and find out the text length, word count and average word length of each summary
- Uni-gram and Multi-gram frequency of words.
- To suggest the best vectorization techniques.
- Use of classification models
- Use of advanced models like transformers, BERT etc.

Significance of the Study

Many patients do not follow instructions as per the health care prescriptions due to various reasons such as understanding the directions, forgetfulness, unpleasant conditions etc. Sticking to a medical routine means sticking to the medicine as prescribed.

It is of prime importance that should focus on improving the health system. This paper intended to create a system that can recommend the correct medicine prescribed by specialties based on the transcriptions.

Scope of the Study / Challenges Involved

The study will include:

i) Medical Transcription
ii) Medical Specialties and its encoded format

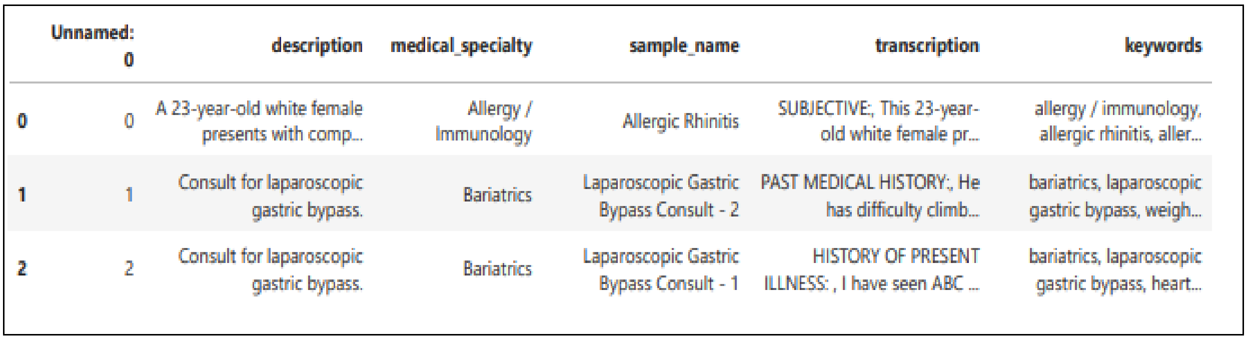
Challenges involved:
Medical data is extremely hard to find due to HIPAA privacy regulations. This dataset offers a solution by providing medical transcription samples. This data was collected from mtsamples.com

V. RESEARCH METHODOLOGY

Please find below the required methodology to be performed for achieving the aims and objectives:

A. Understanding the Data:

Mentioned below are the details of the sample data collected based on the problem statement.



	Unnamed: 0	description	medical_specialty	sample_name	transcription	keywords
0	0	A 23-year-old white female presents with comp...	Allergy / Immunology	Allergic Rhinitis	SUBJECTIVE:, This 23-year-old white female pr...	allergy / immunology, allergic rhinitis, aller...
1	1	Consult for laparoscopic gastric bypass.	Bariatrics	Laparoscopic Gastric Bypass Consult - 2	PAST MEDICAL HISTORY:, He has difficulty climb...	bariatrics, laparoscopic gastric bypass, weigh...
2	2	Consult for laparoscopic gastric bypass.	Bariatrics	Laparoscopic Gastric Bypass Consult - 1	HISTORY OF PRESENT ILLNESS: , I have seen ABC ...	bariatrics, laparoscopic gastric bypass, heart...

The dataset collected has been classified into two categories:

Train Dataset – 2999 rows of train data with transcription and medical specialty to train the model
Test Dataset – 2000 rows of test data with transcription and medical specialty to predict.

The variables described are mentioned in Table 2:

Variable Name	Variable Description	Label
Transcription	Medical Transcript	Feature
medical_specialty	Specialty in the medical department	Target

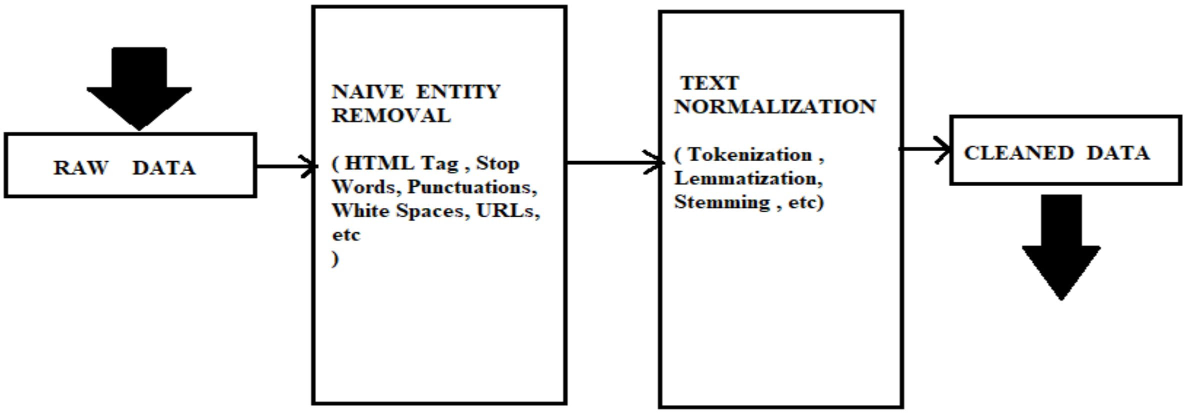
Table 2: Description of the variables

B. Data Pre-Processing:

This is the most important step among all the methodologies used. In these steps, the summary of the data has to be cleaned and effective for the model. The step used includes the removal of HTML tags, the removal of accented character, and removing slash with spaces and punctuations. The text data need to be converted to lowercase and the stop words should also be removed.

Once the above steps are completed successfully, they should be stemmed and lemmatized. This is the process that can change the derived words to their root word which helps in maintaining the information.

The entire steps are described in figure 2 below:


RAW DATA
NAIVE ENTITY REMOVAL
(HTML Tag , Stop Words, Punctuations, White Spaces, URLs, etc)
TEXT NORMALIZATION
(Tokenization , Lemmatization, Stemming , etc)
CLEANED DATA

TEXT CLEANING PIPELINE

Figure 2: Data Cleaning

C. Adding more features relevant to the target column:

It is the best-customized step used for this problem. Using the original features, we have derived a few more features that add more relevancy to the target column. This includes the total length of the summary, word counts of the summary and average length of the summary.

Also based on the Text data, we can use the Sentiment Intensity Analyzer to create columns as polarity, positive, neutral and negative.

All these columns will be effective in improving the accuracy of the model when being trained and evaluated.

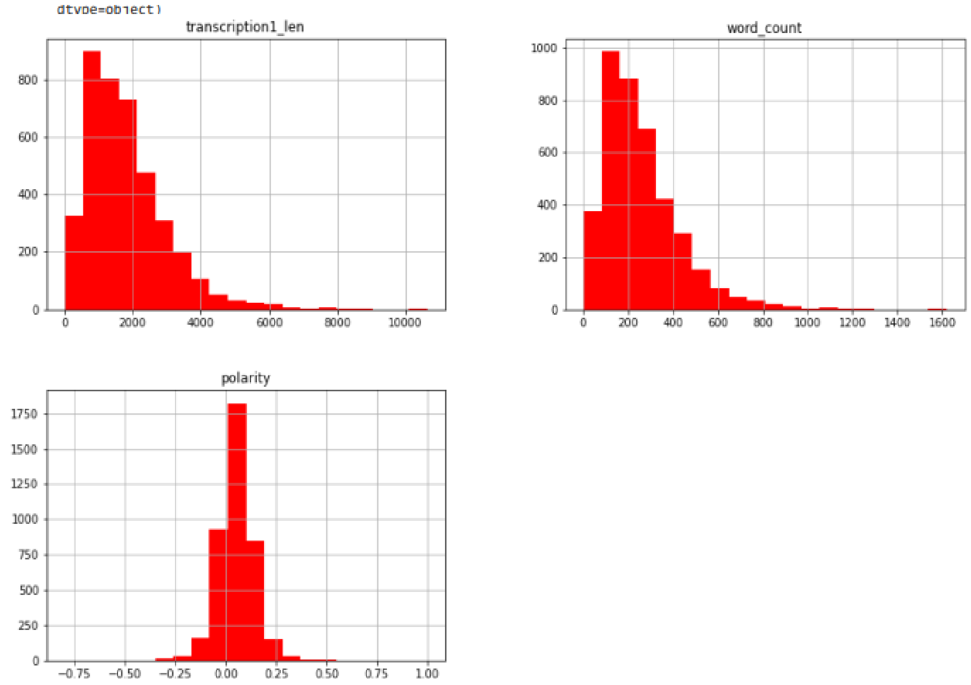


Figure 3

D. Data / Text Visualization:

Visualization of the frequency of words in the text is the immediate step before using the vectorizer. It is to be noted that vectorization can be performed based on count or frequency. Here we focus on the frequency of the words.

The below-mentioned figures (figure 3 and figure 4) represent the bi and tri-gram frequency of words within the text.

Bi-gram Frequency (Two words)

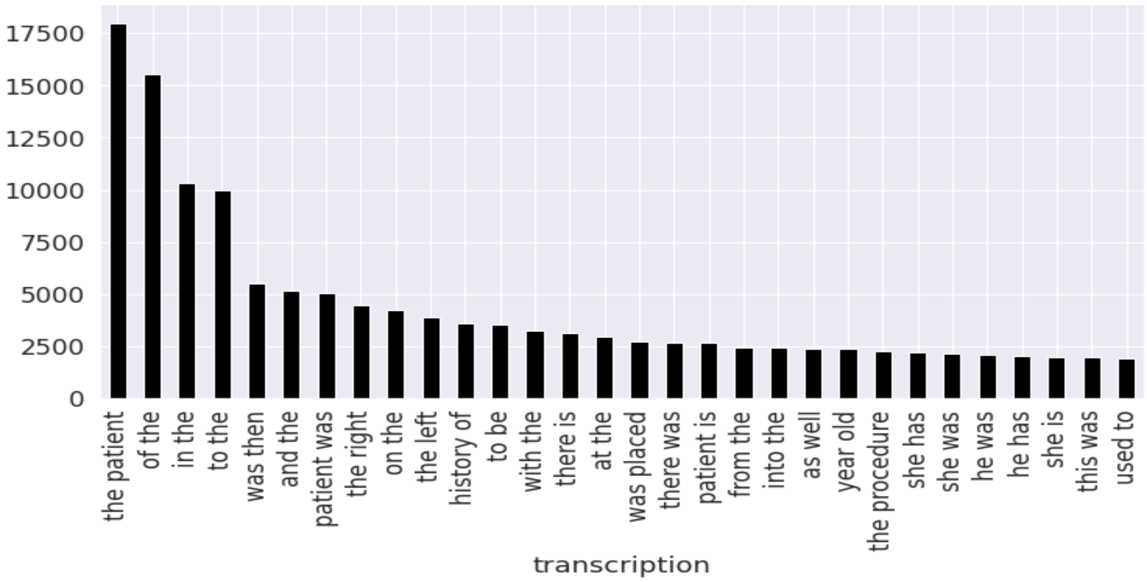


Figure 4

Tri-gram Frequency (Three Words)

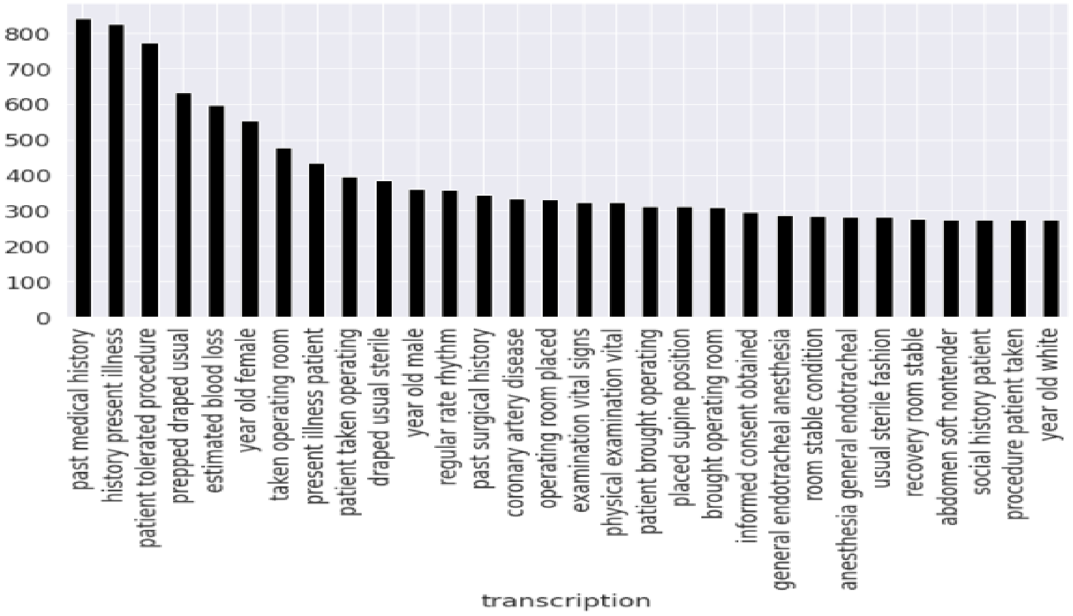


Figure 5

Finally, we are encoding the target variable and then we will use the horizontal stack to combine all the feature variables into a dense format matrix. We should also make sure that the dimensions should be within range to be processed through the classification algorithms.

E. Data Modelling:

Before the model selection, we should be sure that only the dense matrix is considered.

The accuracy of the model should be emphasized to perform a comparative study on the supervised algorithms for this classification problem.

The final step includes optimization of the machine learning models using hyper-parameter tuning to improve the accuracy of the model. The sample parameters used for optimization depend on the selected model. The steps are explained in Figure 6.

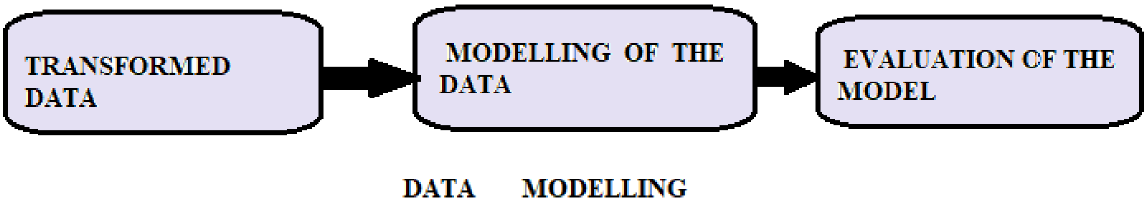
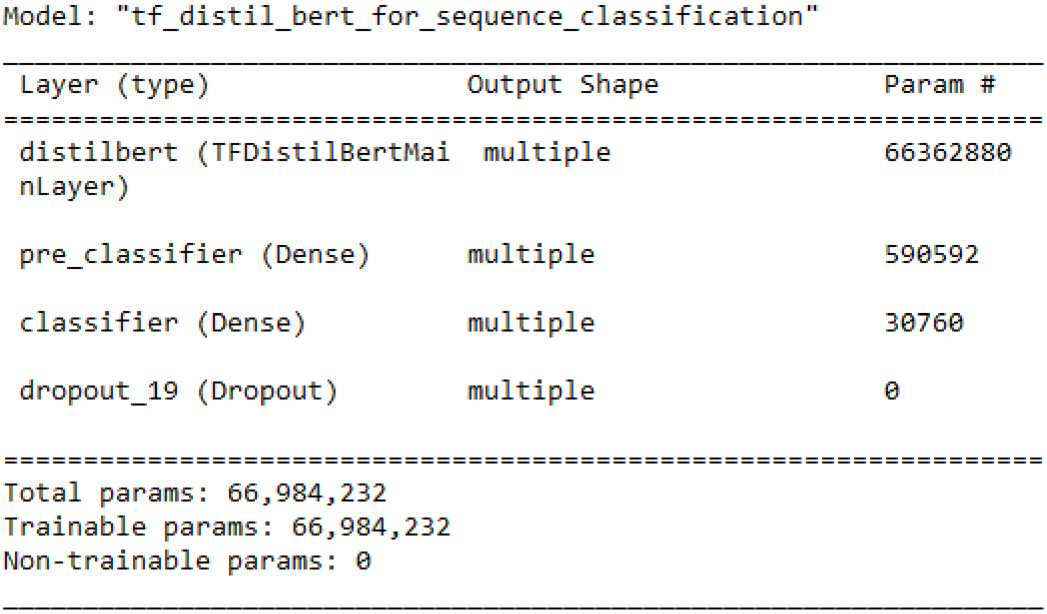


Figure 6

F. Use of BERT

BERT is a deep transformer architecture that uses joint conditioning by combining left and right conditioning within a sentence. It is a kind of encoder composed of a stack of N=4 identical layers. The first one is a multi-head self-attention mechanism, and the second one is a simple position-wise fully connected feedforward network. Mentioned below is the summary of the model used:



```
Model: "tf_distil_bert_for_sequence_classification"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 distilbert (TFDistilBertMai  multiple                 66362880
 nLayer)

 pre_classifier (Dense)      multiple                  590592

 classifier (Dense)          multiple                  30760

 dropout_19 (Dropout)        multiple                  0

=================================================================
Total params: 66,984,232
Trainable params: 66,984,232
Non-trainable params: 0
_________________________________________________________________
```

Figure 7: Summary of the model

Result Obtained:

The final result obtained from the dataset are mentioned below:

Medical Specialty	Prediction
Surgery	195
Consult - History and Phy.	113
Cardiovascular / Pulmonary	65
Radiology	50
Neurology	36
Obstetrics / Gynecology	35
Gastroenterology	33
Ophthalmology	32
Orthopaedic	28
SOAP / Chart / Progress Notes	27
Paediatrics - Neonatal	17
Nephrology	17
Urology	15
Pain Management	13
General Medicine	11
Haematology - Oncology	11
Discharge Summary	10
Office Notes	9
ENT - Otolaryngology	9
Neurosurgery	6
Psychiatry / Psychology	6
Podiatry	5
Emergency Room Reports	4
Dentistry	4
Sleep Medicine	4
Bariatrics	3
Rheumatology	2

Cosmetic / Plastic Surgery	2
Endocrinology	2
IME-QME-Work Comp etc.	2
Physical Medicine - Rehab	2
Lab Medicine - Pathology	1
Letters	1
Dermatology	1
Speech - Language	1

Table 3: Predicted Results

The visual representation of the result predicted is mentioned below:

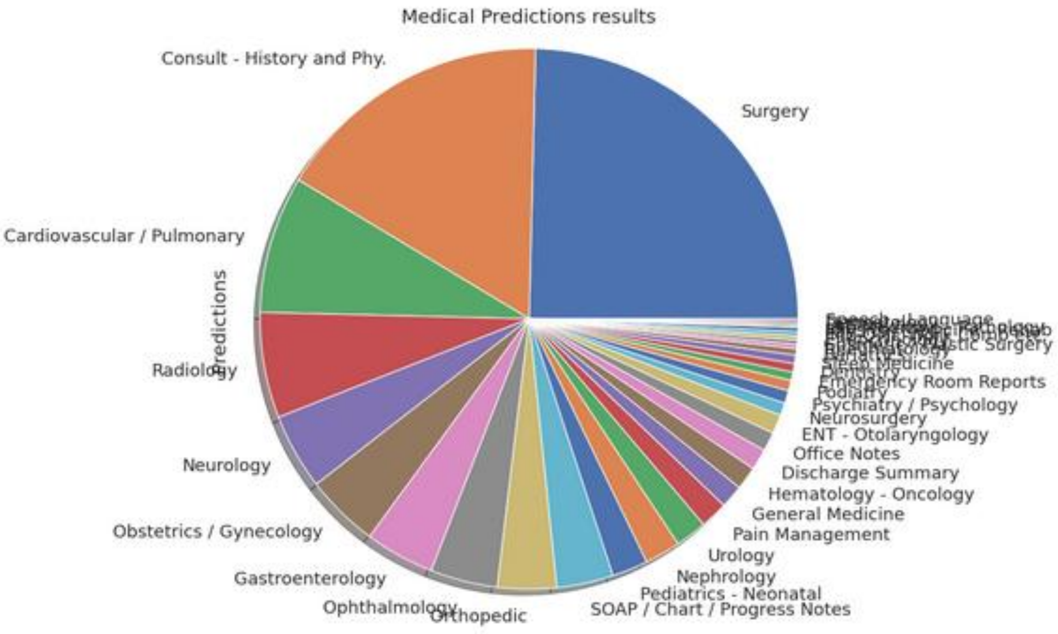


Figure 8: Pie-Chart representation of the result predicted